DIAS - STP - 94 - 34

Queue-lengths and departures at a single-server, multi-class queue *

Neil O'Connell
Dept of Statistics
Trinity College Dublin

November, 1994

In this short note we describe the joint asymptotic tail behaviour of queue-lengths in a single-server, multi-class queue and deduce, in the stationary case, the large deviation properties of the stationary departure processes. The arrivals and service processes need not be independent, and the service may be prioritised. The results may be regarded as a unified generalisation of results in [2, 5, 7, 9]: Chang and Zajic [2] derive the LDP for the stationary departure process from a queue with one arrivals process and stochastic service rate; de Veciana and Kesidis [5] describe the tail-asymptotics for the first queue-length in a multi-class queue with 'generalised processor sharing' (this is a priority service policy); Dupuis and Ellis consider the special case of Markovian inputs; and in [9] the LDP for the departures from an initially empty multi-class system is described.

Queue-lengths

Consider a single-server queue with two arrival processes X_n^1 and X_n^2 and stochastic service capacity C_n shared equally between the two arrival process; for each integer time n, X_n^1 and X_n^2 denote the respective amounts of work arriving at the queue of type 1 and 2, say, and C_n denotes the amount of work that can be serviced (C_n is shared equally between work of type 1 and 2 in the queue, unless one or other amounts to less that $C_n/2$, in which case left-over capacity is given to the other type). Starting with an empty system, the respective queue-lengths at time n are defined recursively by the equations

$$\begin{aligned} Q_n^1 &= (Q_{n-1}^1 + X_n^1 - \max(C_n - Q_{n-1}^2, C_n/2))^+ \\ Q_n^2 &= (Q_{n-1}^2 + X_n^2 - \max(C_n - Q_{n-1}^1, C_n/2))^+ \end{aligned}$$

*Part of this work was carried out while the author was at the Dublin Institute for Advanced Studies, supported by grants from EOLAS and Mentec Computer Systems Limited, under the Higher Education-Industry Cooperation Scheme

with $Q_0^1 = Q_0^2 = 0$. For each n define a path $S_n : [0,1] \to \mathbb{R}^3$ by

$$S_n(t) = \left(\frac{1}{n} \sum_{k=1}^{[nt]} X_k^1, \frac{1}{n} \sum_{k=1}^{[nt]} X_k^2, \frac{1}{n} \sum_{k=1}^{[nt]} C_k \right), \tag{1}$$

and for $\lambda \in \mathbb{R}^3$ set

$$\Lambda(\lambda) = \lim_{n \to \infty} \frac{1}{n} \log E e^{\lambda \cdot S_n(1)}, \tag{2}$$

whenever this limit exists. Write Λ^* for the Legendre transform of Λ.

Denote by $D([0,1], \mathbb{R}^d)$ the space of càdlàg paths (right continuous paths having left limits) in $\mathbb{R}^d$, equipped with the uniform topology, and by $\mathcal{A}C_0^d$ the set of those paths that are absolutely continuous and start at zero.

Theorem 1 *Suppose that Λ^* is a good convex rate function on $\mathbb{R}^3$ and the sequence of paths S_n satisfies the LDP on $D([0,1], \mathbb{R}^3)$ with rate function given by*

$$I(\varphi) = \begin{cases} \int_0^1 \Lambda^*(\dot{\varphi}) dt & \varphi \in \mathcal{A}C_0^3, \\ \infty & \textit{otherwise.} \end{cases} \tag{3}$$

Then for each $\tau > 0$,

$$\lim_{n \to \infty} \frac{1}{n} \log P(Q_{n\tau}^1 \geq a_1 n, Q_{n\tau}^2 \geq a_2 n) = -J_1(a_1, a_2), \tag{4}$$

where

$$J(a_1, a_2) = \inf\{\beta \Lambda^*(x_1, x_2, c) : \; 0 \leq \beta \leq \tau, \; \beta(x_i - c/2) \geq a_i\}. \tag{5}$$

Remark. This was proved in [7] in the case of Markovian fluid sources.

Proof. Without loss of generality we can assume $\tau = 1$. Define a mapping $\Pi : \mathcal{A}C_0^3 \to \mathcal{A}C_0^2$ by letting $q = \Pi(\varphi)$ be the solution to the pair of integral equations ($i \neq j$)

$$q_i(t) = \int_0^t \left[\dot{\varphi}_i(s) - \left[\frac{1}{2} \dot{\varphi}_3 1_{(q_j > 0)} + (\dot{\varphi}_3 - \dot{\varphi}_2) 1_{(q_j = 0)} \right] 1_{(q_i > 0)} \right] ds. \tag{6}$$

For each t, set

$$\tilde{S}_n(t) = S_n(t) + \left(t - \frac{[nt]}{n} \right) \left(S_n \left(\frac{[nt]+1}{n} \right) - S_n \left(\frac{[nt]}{n} \right) \right). \tag{7}$$

Then $\tilde{S}_n \in \mathcal{A}C_0^3$, and

$$\Pi(\tilde{S}_n)(1) = \left(\frac{1}{n} Q_{n\tau}^1, \frac{1}{n} Q_{n\tau}^2 \right). \tag{8}$$

Moreover [4], $\tilde{S}_n$ and S_n are exponentially equivalent, so the sequence $\tilde{S}_n$ also satisfies the LDP on $D([0,1],\mathbb{R}^3)$ with rate function given by I.

For $a_1, a_2 \geq 0$ set

$$B(a_1,a_2) = \left\{\varphi \in \mathcal{A}C_0^3 : \ \Pi_i(\varphi) \geq a_i, \ i = 1,2\right\}. \tag{9}$$

Then, since $B(a_1,a_2)$ is closed, we have by hypothesis

$$\begin{aligned} -\inf_{\varphi\in B(a_1,a_2)^\circ} I(\varphi) &\leq \liminf_n \frac{1}{n}\log P(Q^1_{n\tau} \geq a_1 n, Q^2_{n\tau} \geq a_2 n) \\ &\leq \limsup_n \frac{1}{n}\log P(Q^1_{n\tau} \geq a_1 n, Q^2_{n\tau} \geq a_2 n) \\ &\leq -\inf_{\varphi\in B(a_1,a_2)} I(\varphi). \end{aligned}$$

We will now show (and the statement of the theorem will follow) that for each $\varphi \in \mathcal{A}C_0^3$, there exists a path $\psi \in \mathcal{A}C_0^3$ such that

(i) $I(\psi) \leq I(\varphi)$;

(ii) $\varphi \in B(a_1,a_2) \Rightarrow \psi \in B(a_1,a_2)$;

(iii) $\dot\psi$ is constant on the intervals $[0,\beta)$ and $[\beta,1]$, for some $0 \leq \beta \leq 1$;

(iv) if φ lies in the interior of $B(a_1,a_2)$ then so does ψ.

The path is constructed as follows. Set $\psi(0) = 0$;

$$\dot\psi_i(s) = \begin{cases} \mu_i & 0 \leq s < \beta_{(1)}, \\ \frac{1}{1-\beta_{(1)}}[\varphi_i(1) - \varphi_i(\beta_i)] & \beta_{(1)} \leq s \leq 1, \end{cases} \tag{10}$$

for $i = 1,2$, and

$$\dot\psi_3(s) = \begin{cases} \mu_3 & 0 \leq s < \beta_{(1)}, \\ \frac{1}{1-\beta_{(1)}}\left[(\beta_{(2)} - \beta_{(1)})(\varphi_3(\beta_{(2)}) - \varphi_3(\beta_{(1)})) \right. & \\ \qquad \left. +(1-\beta_{(2)})(\varphi_3(1) - \varphi_3(\beta_{(2)}))\right] & \beta_{(1)} \leq s \leq 1, \end{cases} \tag{11}$$

where

$$\beta_i = \sup\{t : \ \Pi_i(\varphi)(t) = 0\}, \tag{12}$$

$$\beta_{(1)} = \beta_1 \wedge \beta_2, \qquad \beta_{(2)} = \beta_1 \vee \beta_2, \tag{13}$$

and

$$\mu_i = \frac{\partial}{\partial x_i}\Lambda(x_1,x_2,x_3)|_{x_1=x_2=x_3=0}. \tag{14}$$

Claim (i) follows from the convexity of Λ^* and Jensen's inequality. To check (ii) we can, without loss of generality, assume that $\beta_1 \leq \beta_2$. Set $q = \Pi(\varphi)$. Then

$$\begin{aligned} q_1(1) &= \int_{\beta_1}^{1} \dot\varphi_1 - \frac{1}{2}\dot\varphi_3 1_{(q_2>0)} - (\dot\varphi_3 - \dot\varphi_2)1_{(q_2=0)} \\ &\leq \varphi_1(1) - \varphi_1(\beta_1) - \frac{1}{2}[\varphi_3(1) - \varphi_3(\beta_1)] \\ &= \Pi_1(\psi)(1) \end{aligned}$$

(the inequality follows from the fact that $\dot\varphi_2 \leq \dot\varphi_3/2$ whenever $q_2 = 0$), and

$$q_2(1) = \int_{\beta_2}^{1} \dot\varphi_2 - \dot\varphi_3/2 = \Pi_2(\psi)(1);$$

this implies (ii). □

If X^i and C are stationary sequences and $E(X_1^1 + X_1^2) < EC_1$, then the sequence (Q_n^1, Q_n^2) has a stationary distribution (Q^1, Q^2) and we can deduce (informally) from Theorem 1 that

$$\lim_{n\to\infty} \frac{1}{n} \log P(Q^1 \geq a_1 n, Q^2 \geq a_2 n) = -J(a_1, a_2), \tag{15}$$

where

$$J(a_1, a_2) = \inf\{\beta\Lambda^*(x_1, x_2, c) : \beta \geq 0,\ \beta(x_i - c/2) \geq a_i\}. \tag{16}$$

The proof of Theorem 1 can be modified to yield the following generalisation, where we have l arrival processes $X^1, \ldots, X^l$ and a 'generalised processor sharing' priority service scheme with weights $p_1, \ldots, p_l$ summing to 1: these systems are formally defined in [5], where the asymptotic tails of the first queue-length are characterised. In this case we put

$$S_n(t) = \left(\frac{1}{n}\sum_{k=1}^{[nt]} X_k^1, \ldots, \frac{1}{n}\sum_{k=1}^{[nt]} X_k^l, \frac{1}{n}\sum_{k=1}^{[nt]} C_k \right), \tag{17}$$

and for $\lambda \in \mathbb{R}^{l+1}$ define

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n} \log Ee^{\lambda\cdot S_n(1)}, \tag{18}$$

whenever this limit exists, writing Λ^* for the Legendre transform of Λ. Denote by $Q_n^1, \ldots, Q_n^l$ the respective queue-lengths at time n.

Theorem 2 *Suppose that Λ^* is a good convex rate function on $\mathbb{R}^{l+1}$ and the sequence of paths S_n satisfies the LDP on $D([0,1], \mathbb{R}^{l+1})$ with rate function given by*

$$I(\varphi) = \begin{cases} \int_0^1 \Lambda^*(\dot\varphi)dt & \varphi \in \mathcal{AC}_0^{l+1}, \\ \infty & \textit{otherwise}. \end{cases} \tag{19}$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log P(Q^1_{n\tau} \geq a_1 n, \ldots, Q^l_{n\tau} \geq a_l n)$$
$$= \; -\inf\{\beta\Lambda^*(x_1, \ldots, x_l, c) : \; 0 \leq \beta \leq \tau, \; \beta(x_i - p_i c) \geq a_i\}.$$

If the process $(Q^1_n, \ldots, Q^l_n)$ has a stationary distribution, then, assuming the interchange of limits $n \to \infty$ and $\tau \to \infty$ is justified, the tails of this distribution satisfy

$$\lim_{n\to\infty} \frac{1}{n} \log P(Q^1 \geq a_1 n, \ldots, Q^l \geq a_l n)$$
$$= \; -\inf\{\beta\Lambda^*(x_1, \ldots, x_l, c) : \; \beta \geq 0, \; \beta(x_i - p_i c) \geq a_i\}.$$

Departures

As an application of the above, we extend (without formal proof) a result of [9] to the stationary case. In [9] the LDP for the departure processes was described in the case where the queue is assumed to be initially empty; now that we know the tail behaviour of the distribution of queue-lengths in equilibrium, we can write down the implied LDP for the departures from a queue which is initially in equilibrium.

This has been done by Chang and Zajic [2] in the case of one arrivals process; they made the important observation that when the service rate is stochastic, the stationary departure process need not have the same large deviation properties as the departure process from an initially empty queue. This is because a large deviation in the departure process can be encouraged by starting with a very long queue. Formally, their result can be stated as follows. Let $\Lambda^*(x, c) = \Lambda^*_a(x) + \Lambda^*_s(c)$ be the rate function corresponding to the arrivals and service processes (they are assumed to be independent). Then [1, 3, 6, 8] the tails of the stationary queue length distribution satisfy

$$\lim_{n\to\infty} \frac{1}{n} P(Q \geq qn) = -\delta q, \tag{20}$$

where

$$\delta = \inf_{x,c} x^{-1}[\Lambda^*_a(x + c) + \Lambda^*_s(c)], \tag{21}$$

and [9] the cumulative departures from an empty queue satisfy the LDP with rate function

$$\Lambda^*_D(z) = \Lambda^*_a(z) + \Lambda^*_s(z \vee \Lambda'_s(0)). \tag{22}$$

Chang and Zajic [2] show that, under additional mixing hypotheses, the stationary departures satisfy the LDP with rate function given by

$$\tilde{\Lambda}^*_D(z) = \inf_q[\delta q + \Lambda^*_D(z - q)]. \tag{23}$$

The mixing hypotheses are required because the queue length at time zero is not independent of subsequent service and arrivals.

Now suppose there are two arrivals processes (types 1 and 2): denote by D_n^i the cumulative departures of type i upto time n. If the queue is initially empty ($Q_0^1 = Q_0^2 = 0$), then a slight modification of the proof of Theorem 3.1 in [9] yields:

Theorem 3 *Suppose the sequence of paths S_n satisfies the LDP on $D([0,1], \mathbb{R}^3)$ with rate function given by*

$$I(\varphi) = \begin{cases} \int_0^1 \Lambda^*(\dot{\varphi})dt & \varphi \in \mathcal{A}C_0^3 \textit{ non-decreasing,} \\ \infty & \textit{otherwise.} \end{cases} \tag{24}$$

Then the sequence of paths $(D_{[n\cdot]}^1/n, D_{[n\cdot]}^2/n)$ satisfy the LDP on $D([0,1], \mathbb{R}^2)$ with rate function given by

$$I(\varphi) = \begin{cases} \int_0^1 \Lambda_D^*(\dot{\varphi})dt & \varphi \in \mathcal{A}C_0^2 \textit{ non-decreasing,} \\ \infty & \textit{otherwise,} \end{cases} \tag{25}$$

where

$$\begin{aligned} \Lambda_D^*(z_1, z_2) = \inf\{\beta(x,y,c)\Lambda^*(x,y,c) + (1-\beta)\Lambda^*(\mu_1,\mu_2,c) : \\ x\beta(x,y,c) = z_1,\ y\beta(x,y,c) = z_2\}. \end{aligned}$$

As in the case of one arrivals process, we can combine this with Theorem 1 to get that, under suitable mixing hypotheses, the stationary departures satisfy the LDP with rate function givem by

$$I(\varphi) = \begin{cases} \int_0^1 \tilde{\Lambda}_D^*(\dot{\varphi})dt & \varphi \in \mathcal{A}C_0^2 \text{ non-decreasing,} \\ \infty & \text{otherwise,} \end{cases} \tag{26}$$

where

$$\tilde{\Lambda}_D^*(z_1, z_2) = \inf\{J(a_1, a_2) + \Lambda_D^*(z_1 - a_1, z_2 - a_2) :\ a_i \le z_i\}. \tag{27}$$

This can of course be generalised to an arbitrary number of inputs with a priority service scheme.

References

[1] Cheng-Shang Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. on Automatic Control* 39:913–931, 1994.

[2] Cheng-Shang Chang and Tim Zajic. Effective bandwidths of departure processes from queues with time varying capacities. INFOCOM 1995.

[3] G. de Veciana, C. Courcoubetis and J. Walrand. Decoupling bandwidths for networks: a decomposition approach to resource management. Memorandum No. UCB/ERL M93/50, University of California.

[4] Amir Dembo and Tim Zajic. Large deviations: from empirical mean and measure to partial sums process. Preprint.

[5] G. de Veciana and G. Kesidis. Bandwidth allocation for multiple qualities of service using generalised processor sharing. Preprint.

[6] N.G. Duffield and Neil O'Connell. Large deviations and overflow probabilities for the general single server queue, with applications. *Proc. Camb. Phil. Soc.*, to appear.

[7] Paul Dupuis and Richard S. Ellis. The large deviation principle for a general class of queueing systems, I. *Trans. Amer. Math. Soc.*, to appear.

[8] Peter W. Glynn and Ward Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.*, to appear.

[9] Neil O'Connell. Large deviations in queueing networks. DIAS technical report DIAS-APG-9413, 1994.